D... COMME DISTOGRAMME

Didier JOSSELIN *: THEMA UPRESA 6049 du CNRS, Université de Franche-Comté*
Email ***:*** ***didier.josselin@univ-fcomte.fr***

RESUME. *Le Distogramme est un outil dédié à l'analyse des discontinuités statistiques et spatiales, grâce au lien dynamique entre une représentation cartographique et des histogrammes de variables descriptives. Il permet de modifier interactivement le nombre ou les limites de classes, de discrétiser, transformer et croiser les variables.*

ABSTRACT *: : The Distogram is an exploratory tool provided to analyse statistical and geographical discontinuities, by a dynamic link between a map and histograms related to different variables. It provides to modify interactively the classes number and limites, to trim, transform and cross the available variables*

MOTS CLEFS *: : Analyse exploratoire des données géographiques, discrétisation, discontinuités, lien dynamique.*

KEYS WORDS *: Exploratory Spatial Data Analysis, discretization, discontinuities, dynamic link*

Le monde qui nous entoure n'est pas une juxtaposition aléatoire d'objets géographiques sans interaction. Il ne forme pas un ensemble homogène, isomorphe. Qu'il s'agisse d'agglomérations urbaines, de répartitions d'assolements agricoles, l'espace géographique est découpé, organisé par un ensemble de processus de différenciation socio-spatiale. S'attacher à détecter les franges de processus d'évolution, les zones de discontinuité, l'existence d'agrégats ou d'auto corrélation spatiale, constitue l'une des tâches du géographe. Ce sont ces raisons qui ont motivé initialement le développement du Distogramme.

Fondamentalement, le Distogramme ne fait que reprendre un ensemble d'outils et de concepts couramment utilisés en géographie. Il repose sur des méthodes de discrétisation éprouvées (CAUVIN, REYMOND, SERRADJ, 1987) et d'animations cartographiques. Sa particularité est qu'il les regroupe en un ensemble cohérent et les met au service de l'expert dans un processus interactif. Il ne s'agit aucunement d'un outil de cartographie au sens strict, mais plutôt d'un outil d'analyse exploratoire, pour visualiser et identifier inter activement tous les individus afin de les associer ou les regrouper.

Pratiquement, le Distogramme est issu :

- du besoin d'une meilleure interactivité entre l'histogramme d'une variable et la représentation cartographique des individus décrits,
- de la difficulté de discrétiser correctement certaines variables, même transformées,
- de la nécessité de prendre en compte la dimension spatiale dans la construction d'un histogramme « optimal »,
- de l'insuffisance des méthodes de discrétisation automatique fournies par les logiciels de cartographie ou les SIG.

L'objectif du Distogramme est d' *« Analyser dynamiquement les distributions spatiales et statistiques pour permettre une meilleure appréhension des objets géographiques et des relations qu'ils entretiennent (fonctionnelles, structurelles, statistiques, spatiales, topologiques…) ».*

1. D… comme Double

Le Distogramme, développé dans l'environnement statistique d'Xlisp-Stat (TIERNEY, 1990) associe **deux outils** complémentaires.

Le premier outil est *la carte*. Nul n'est besoin ici de rappeler en détails ses vertus (BRUNET, 1987), la seule fonctionnalité de cartographie d'un phénomène ou de résidus constituant déjà un pas important dans l'analyse spatiale. Au sein du Distogramme, elle est réduite à sa plus simple expression, puisque la plupart des règles de sémiologie ne sont pas intégrées. L'information y est géocodée et des couleurs permettent d'identifier les individus regroupés en classes. Trois niveaux de complexité peuvent alors être appréhendés :

- les localisations spatiales absolues des entités géographiques,
- les localisations relatives entre ces entités (distances, dont topologiques, dispersions spatiales, etc.),
- les structures et les formes spatiales qui en découlent (agrégats, objets géographiques composites, etc.) (JOSSELIN, 1999).

Le second outil est *l'histogramme*. Qui n'a pas appréhendé un problème en réalisant, en première étape d'une analyse statistique, une belle distribution d'une variable nominale, ordinale ou cardinale ? La distribution statistique reste un outil puissant de simplification de l'information, même si elle est parfois discutée. En effet, elle est moins robuste que d'autres représentations, tel que le Dotplot (1 point = 1 individu, empilés quand ils sont proches), puisqu'elle ne fournit pas une représentation exhaustive des individus, mais les regroupe au sein des classes (FLOCH et al., 1998). Le choix de l'histogramme réside dans sa capacité intrinsèque à la discrétisation de variables et à la recherche de discontinuités. Figé, il peut être utilisé d'au moins trois façons :

- synthétiquement, en calculant des indicateurs centraux, telle que la médiane,
- globalement, en analysant et en décrivant sa forme (nombre de modes et leurs valeurs, calcul des coefficients d'aplatissement et d'asymétrie, etc.),

– localement, en observant la position des individus dans la distribution (fréquences par classe, positions relatives et répartitions des individus dans les classes).

Toutefois, dans un processus d'analyse spatiale, il est légitime de se poser quelques questions quant à l'utilisation séparée de ces deux outils. La cartographie met-elle bien en évidence les traits du phénomène que je cherche à analyser ? La variable quantitative est-elle correctement discrétisée pour révéler une répartition statistique ou spatiale particulière de mes individus ? Quelle peut être l'influence d'une modification de classe ? Les discontinuités spatiales dont je soupçonne l'existence apparaissent-elles ?

Ces questions font référence à **deux concepts** sur lesquels repose le Distogramme.

D'une part, analyser séparément la carte et l'histogramme reste un moyen limité pour appréhender l'espace dans sa continuité et sa diversité. Faire interagir ces deux outils complémentaires dans un processus exploratoire donne une dimension nouvelle à l'analyse. Après les *distibutions statistiques* et *spatiales*, le *lien dynamique* est le troisième mot clé (HASLETT et al., 1991).

D'autre part, il semble nécessaire de pouvoir modifier empiriquement la structure des classes dans l'histogramme, et ce de la manière la plus conviviale et la plus rapide possible. Cela constitue une quatrième façon d'utiliser l'histogramme : dynamiquement, en simulant l'impact de modification de discrétisation d'une variable sur une cartographie.

Tout ou partie de ces deux concepts est déjà mis en œuvre dans certains Systèmes d'Information Géographique (MacMap ou Géoconcept, par son module «Thématique»), certains logiciels de cartographie automatique (Cartes&Données, par exemple) ou dans les logiciels du domaine de l'analyse exploratoire des données : XlispStat par Tierney (1990), Datadesk (Waniez, 1991) et Philexplo (Waniez, 1999).. Certains développements à la confluence entre l'analyse spatiale et la statistique existent également (Livemap par Brunsdon, 1998, ARPEGE [1] et Lav-Stat [2] par Josselin et al., 1999, SpatialStat et SpaceStat).

2. D... comme Dynamique

Le Distogramme lie en permanence une carte et une (ou des) distribution(s) statistique(s) de variable(s) descriptive(s) à cartographier. L'expert peut le faire évoluer en fonction de ses investigations.

On peut sélectionner des individus sur la carte et constater leur répartition dans les classes de l'histogramme (fig. 1). L'inverse est également possible : on choisit graphiquement une ou plusieurs classes, dont les individus sont identifiés en même temps sur la carte (fig. 2). Par ailleurs, modifier la distribution fait basculer des individus d'une classe dans une autre et déclenche la nouvelle cartographie. Ce lien dynamique permanent est un élément clé qui permet à l'expert de valider/invalider immédiatement ses hypothèses de discontinuité des points de vue statistique et spatial.

Nous présentons en exemple l'étude des flux de sportifs dans les communes du périmètre du Schéma Directeur du Grand Besançon, les deux variables étudiées sont :

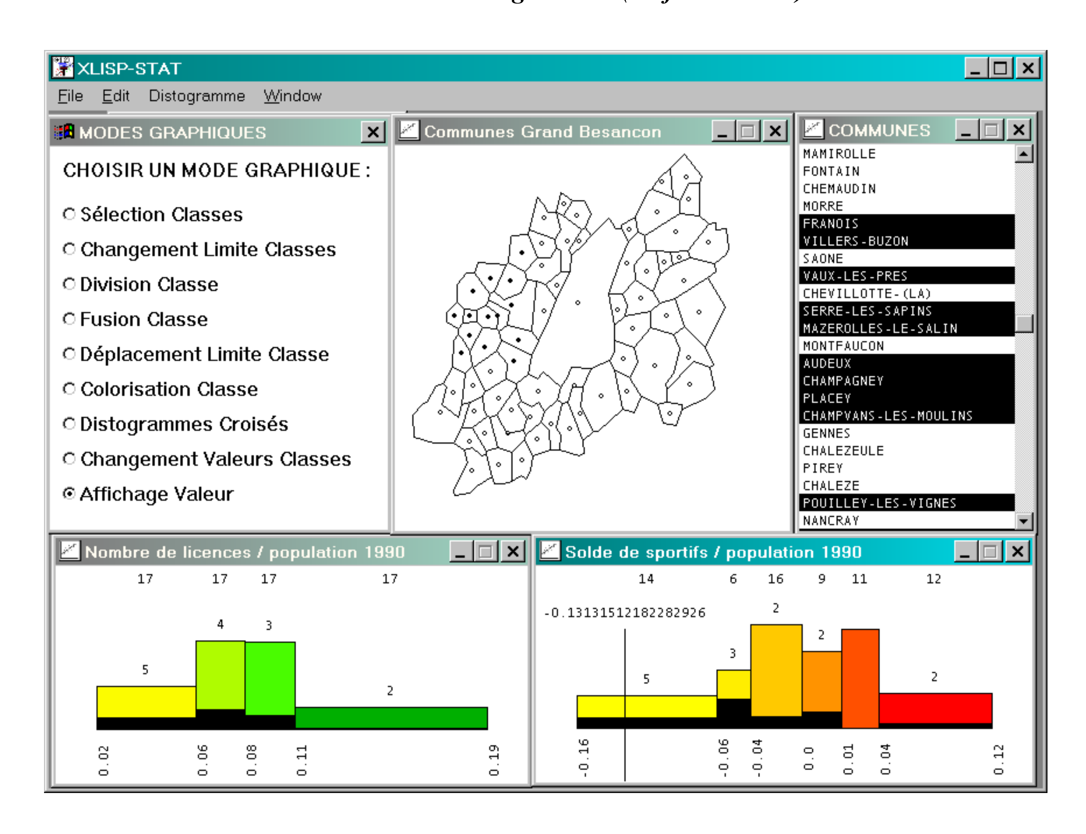
- le solde de sportifs (nombre de licenciés extérieurs - tous sports confondus - venant jouer dans la commune moins les habitants allant pratiquer un sport dans d'autres communes),

- le nombre de licenciés,

[1] Analyses Robustes Pour l'Exploration GEographique

[2] Lien dynamique entre Arc View et XlispStat

Fig.1 : Le Distogramme : sélection de communes sur la carte (centroïdes en noir) et visualisation de leurs distributions dans les deux histogrammes (surfaces noires) et dans la liste des communes



Le lien dynamique entre diverses représentations et aspects d'un même objet d'étude est donc un élément fondamental qui permettra de mettre en phase le processus d'apprentissage et d'investigation de l'expert avec le processus informatique. Nous pensons que sa présence bonifie les analyses dans leur ensemble, grâce à une approche de type systémique.

3. D... comme Discrétisation

Un certain nombre de fonctionnalités du Distogramme concernent directement la discrétisation de variable quantitative (fig. 2).

Parfois, les discrétisations automatiques proposées (classes d'amplitude égale, répartition en quantiles, par exemple) ne rendent pas bien compte des groupes d'individus. Si une classe englobe deux sous-groupes, il peut être opportun de la diviser. A l'opposé, deux classes possédant peu d'effectifs ou considérées par l'expert comme proches sémantiquement peuvent être regroupées. Il peut être utile de réaliser des discrétisations de variables «à façon» en intégrant trois critères complémentaires :

- le critère «de construction» (règle de découpage et nombre de classes préalable),
- le critère «statistique» (bonne répartition et homogénéité des individus dans les classes),
- le critère «sémantique» (évaluation de découpages par l'expert, prise en compte des répartitions spatiales sur la carte associée).

Le troisième critère permet à l'expert de peaufiner sa recherche de découpages, d'agrégats, de gradients ou de discontinuités dans l'espace géographique.

Fig.2 : Le Distogramme: sélection graphique des individus de trois classes d'une distribution et analyse de leur répartition spatiale

4. D... comme Discontinuité

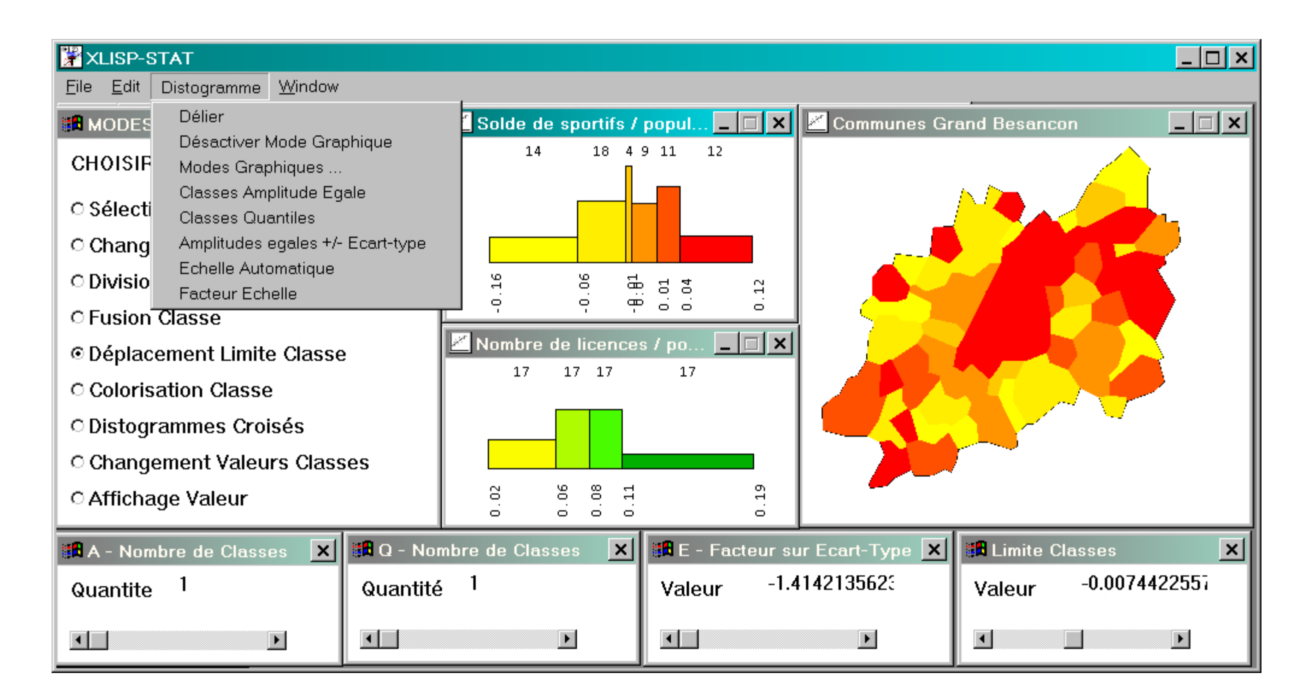
Au-delà des possibilités offertes pour construire des distributions dédiées, le Distogramme propose plusieurs fonctionnalités pour analyser les discontinuités d'une variable quantitative dans l'espace géographique (fig. 3).

La première consiste à modifier, avec un curseur, le nombre de classes de la distribution. La méthode automatique de discrétisation proposée propose un découpage soit par classes d'amplitudes égales, par quantiles, soit par multiples d'écart type centré sur la moyenne. Le fait d'augmenter de manière continue le nombre de classes et d'en constater les modifications sur la carte permet, d'une part, d'évaluer l'effet de la méthode de discrétisation sur la cartographie, d'autre part, de diminuer de plus en plus la résolution dans l'analyse.

Il peut être opportun, dans certains cas, de se focaliser sur la limite entre deux classes, et de la modifier de façon manuelle (on choisit sa nouvelle position) ou graduelle (avec un curseur, on la promène avec un pas défini). Ce procédé est intéressant dans la mesure où il permet de « brosser » les individus en analysant leurs positions respectives sur la carte et dans la distribution.

L'identification des discontinuités ou des structures agrégées peut faire appel à ces approches : il suffit de relâcher la pression sur le curseur dès qu'un nombre important d'individus, parfois proches (géographiquement ou topologique ment), changent de classe. Alors, une quantité d'individus, la position d'une limite de classes sont peut-être discriminants (fig. 3).

Fig.3 : Le Distogramme : un outil interactif de discrétisation par modification du nombre de classes ou de leurs limites

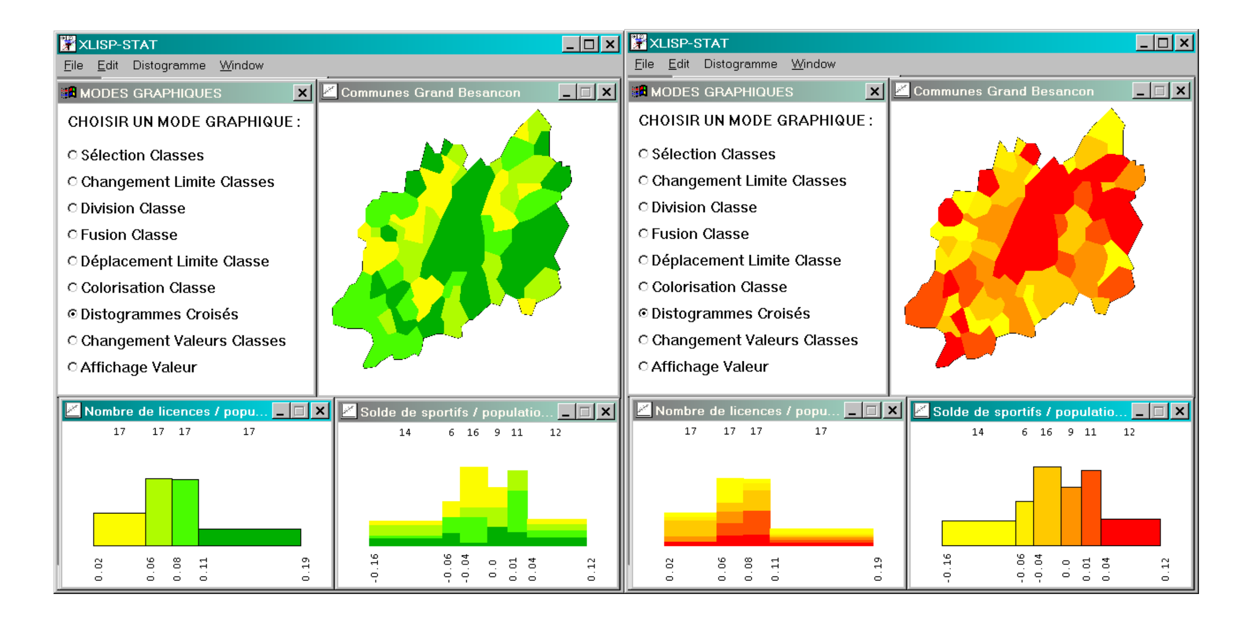


5. D... comme Distributions croisées

Lors d'une même analyse, il est possible d'associer plusieurs variables à une même carte. La discrétisation d'une variable peut également dépendre de la relation statistique qu'elle entretient avec d'autres variables. Le géographe doit pouvoir en tenir compte dans son analyse.

C'est pourquoi nous avons implémenté (et modifié) le croisement des distributions (fig. 4), tel qu'il existe dans certains logiciels d'analyse exploratoire, comme Datadesk ou XlispStat. Derrière cette fonctionnalité, se cache tout simplement la notion de contingence. En effet, visualiser la répartition des individus par classe d'une variable A au sein d'une distribution d'une variable B revient à réaliser un tableau de contingence et à comparer les effectifs réels à des effectifs théoriques. Si l'on constate une équirépartition des classes de A dans B, les deux variables sont indépendantes statistiquement. Des individus d'une classe de A qui occupent la majeure partie d'une classe de B marquent une dépendance statistique positive mais négative si elle couvre une surface plus petite qu'attendu (hypothèse d'indépendance statistique). Un choix manuel des couleurs (RVB) offre la possibilité à l'utilisateur de personnaliser son analyse par l'emploi des couleurs (fig. 1).

Fig.4 : Distributions croisées dans le Distogramme



6. D... comme Distorsion de valeurs

Qui n'a pas été confronté, dans le dépouillement d'enquêtes ou l'analyse de données statistiques, à des distributions «anormales», avec des «pics» sans organisation apparente ? Quelque soit la méthode de discrétisation, l'investigation reste délicate. En découpant par amplitudes de classes égales, on observe des classes vides et des classes pleines. En utilisant les quantiles, l'hétérogénéité des valeurs fait que certaines classes fines peuvent atteindre des sommets en X (les individus sont rares et leurs valeurs s'étalent) ou en Y (les individus sont nombreux et leurs valeurs sont proches). Cela provient de la contrainte de lisibilité imposée par les histogrammes : la surface des bâtonnets est proportionnelle aux effectifs de la classe.

Dans le cadre d'une analyse par Distogramme, cette contrainte est respectée et rend difficile la discrétisation. L'idéal serait de disposer d'individus mieux répartis sur l'amplitude de la variable afin de pouvoir plus facilement les sélectionner et les trier graphiquement. Il peut être alors opportun de modifier la structure ou le contenu de la distribution. Trois solutions à ce problème sont possibles.

On peut, tout d'abord, modifier les axes X et / ou Y. Mais réaliser une simple homothétie ne ferait que repousser le problème (changement d'échelle sur 1 ou 2 axe(s)).

La seconde solution est celle retenue par les logiciels d'analyse exploratoire : se focaliser sur un sous-ensemble d'individus et étudier ceux-ci indépendamment des autres. Ce choix ne permet pas une prise en compte globale de la population.

La troisième solution consiste à transformer les valeurs. C'est cette voie que nous proposons de suivre dans le Distogramme. L'idée est de rapprocher ou d'étaler certaines valeurs parmi lesquelles une limite de classe pourrait apparaître (fig. 5).

L'expérience montre que rares sont les distributions empiriques qui peuvent être facilement transformables (cas, par exemple, des distributions hyperboliques). Pour celles-ci, on peut appliquer une transformation globale des valeurs (par le log, par exemple). Pour les autres, nous proposons d'appliquer une transformation locale par classe. Nous définissons ainsi une fonction de transformation pour tout x d'une classe ; par exemple, fig. 5:

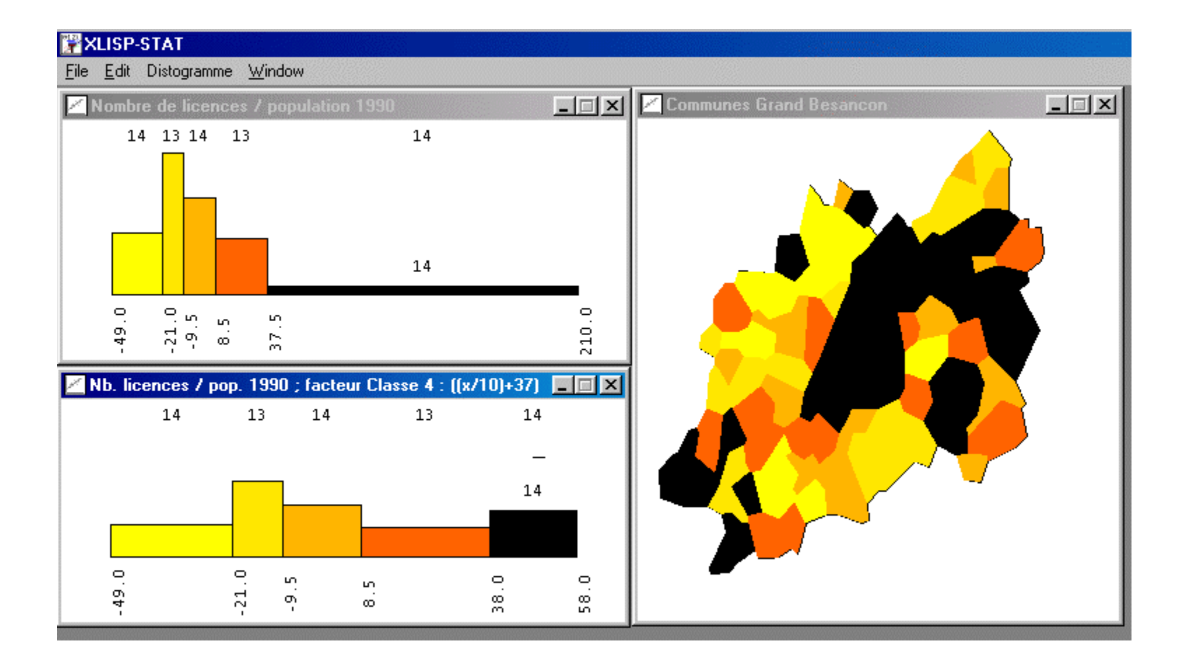
division par 10 et ajout de 37 aux valeurs de la classe 4. [3]. Cette fonction correspond en fait à une «ré-expression» locale. Les fonctions peuvent être différentes dans les classes, si l'on y observe des comportements spécifiques.

En pratique, cette méthode pose de sérieux problèmes de cohérence de l'information : comment comparer les individus transformés avec les autres ? Plusieurs contraintes doivent être respectées :

- l'ordre des valeurs doit rester le même (à des fins de discrétisation ultérieures),
- les valeurs maximales et minimales transformées ne doivent pas excéder les valeurs initiales, afin de rester dans la bonne classe (dans le cas inverse, nous assisterions à un mélange inextricable de valeurs brutes et recalculées),
- l'utilisateur doit bien garder à l'esprit quelles classes ont été transformées et de quelle façon,
- il doit bien repérer quelles sont les limites des classes avant et après transformation,
- lors de la recherche de discontinuités, il doit assumer qu'un déplacement élémentaire de limite de classe n'a pas toujours la même signification selon la classe considérée (notamment pour une limite qui sépare la classe transformée d'une autre).

Cette méthode requiert donc une attention particulière, à cause de la non-linéarité de l'abscisse de la variable étudiée et des effets potentiels des transformations de valeurs par classe. On aboutit en définitive à une nouvelle distribution, plus facilement exploitable graphiquement, mais qui ne peut réellement constituer un document statistique de synthèse diffusable en l'état (fig. 5). Il convient donc de garder la distribution initiale comme référence, et, grâce à la fonctionnalité de croisement dynamique, de vérifier en permanence où se trouvent les individus dans les deux distributions. Le rôle de la nouvelle distribution n'est que de faciliter l'analyse exploratoire des discontinuités spatiales : elle permet une investigation plus précise dans les classes de fortes densités, tout en conservant l'ensemble des individus observés.

Fig. 5 : Transformation locale de variable dans le Distogramme



[3] en LISP : *(defun fontion (x)* (+37 *(/ x 10)))*

7. Conclusion

Actuellement, l'analyse exploratoire des données spatiales (ESDA[4], FOTHERINGHAM et al., 2000) est un domaine de recherche qui se développe fortement. Elle découle directement de l'EDA (Exploratory Data Analysis, TUKEY, 1977, HOAGLIN et al., 1983) qui ne concernait au départ que la statistique. Elle met en avant, entre autres, la démarche empirique et qualitative, la robustesse des outils statistiques employés, l'importance de la prise en compte des individus autant que de la tendance. Le distogramme fait partie de ce courant, qui semble adapté à l'analyse de données multisources, multiscalaires et incertaines, comme le sont souvent les informations géographiques. Dans ce contexte, il améliore la robustesse de l'histogramme par des fonctionnalités variées de discrétisation de variables et facilite la recherche de discontinuités spatiales grâce au lien dynamique entre les diverses représentations statistiques et cartographiques.

BIBLIOGRAPHIE

BRUNET R., 1987, La carte mode d'emploi, Fayard/Reclus, 269 p.

BRUNSDON C., 1998, Exploratory spatial data analysis and local indicators of spatial association with XlispStat, The Statistician, n°47, Part 3, pp. 471-484.

CAUVIN C., REYMOND H., SERRADJ A, 1987., *Discrétisation et représentation cartographique*, Collection Reclus Mode d'Emploi, 116 p. + annexes

FLOCH JM., GRUN-REHOMME M., LADIRAY D., 1998, Exploratory Data Analysis, Cours de 3[ème] année d'ENSAE, 150 p.

FOTHERINGHAM A. S., BRUNSDON C., CHARLTON M., 2000*, Quantitative Geography, Perspectives on Spatial Data Analysis*, SAGE Publications, London, 270 p.

HAOGLIN D., MOSTELLER F., TUKEY J.W., 1983, *Understanding robust and exploratory data analysis*, Wiley Series in probability and mathematical statistics, 447p.

HASLETT J., BRADLEY R., CRAIG P., UNWIN, A., WILLS G., 1991, Dynamics graphics for exploring spatial data with application to locating global and local anomalies in *The American Statistician*, August 1991, vol. 45, N° 3, pp. 235-242

JOSSELIN D., CHATONNAY P., GUERRE L., DANCUO B., 1999, Lien dynamique entre ArcView et Xlisp-Stat (LAVSTAT) : un environnement interactif d'analyse spatiale, Actes de la Conférence Française des Utilisateurs ESRI, 29-30 septembre 1999, Cédérom,

JOSSELIN, 1999, A la recherche d'objets géographiques composites, N° spécial Data Mining Spatial, *Revue Internationale de Géomatique,* pp. 489-505, Vol. 9, 4

TIERNEY L., 1990, *Lisp-Stat, an object oriented environment for statistical computing and dynamic graphics*, John Wiley and Sons, NewYork, 350 p.

TUKEY JW, 1977*, Exploratory data Analysis*, Addison-Wesley.

WANIEZ P., 1991, Analyse exploratoire des données, GIP Reclus, Reclus Mode d'Emploi, n° 17, Montpellier.

WANIEZ P., 1999, La cartographie des données économiques et sociales sur Macintosh et PowerMacintosh avec Philcarto et Philexplo, L'Harmattan, Paris.

[4] Exploratory Spatial Data Analysis